Make Chicago "Safe" Again

Data Mining Project

Luis Veltze
University of Colorado at Booulder
Boulder, Colorado
luis.veltze@colorado.edu

Ryan Close
University of Colorado at Boulder
Boulder, Colorado
ryan.close@colorado.edu

Tyler Mooney
University of Colorado at Boulder
Boulder, Colorado
tyler.mooney@colorado.edu

Garrett Glissmann
University of Colorado at Boulder
Boulder, Colorado
garrett.glissmann@colorado.edu

ABSTRACT

The goal of this project is to find interesting trends using Chicago's crime reports spanning the years 2001 to early 2018. This team will perform a longitudinal survey of crime trends. The goal is to find geographical patterns as well as discover connections to other public health factors and socioeconomic indicators. For example, we could correlate crime with birth rates or access to affordable housing and grocery stores. What the team found was that the overall trend of crime in Chicago since 2001 is downwards. However, there are variations in the amount of crime during the year as well as by type of crime.

ACM Reference format:
Luis Veltze, Ryan Close, Tyler Mooney, and Garrett Glissmann. 2016. Make Chicago "Safe" Again. In *Proceedings of CSCI-4502,*
, Boulder (University of Colorado at Boulder), 8 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the search to find a data set that would be a good candidate for data mining, the team considered datasets about bitcoin and crime. We selected crime data because it was more intuitive to understand the attributes and ultimately mine correlations between similar datasets. Crime data is one of many subjects that is continuously examined and analyzed for trends and patterns. Of course, studying crime has an important social value: to understand how much crime is happening, how it compares in terms of time and geography, and possible reasons that crime rises or falls. The hope is that in understanding the patterns and causes of crime, it might be possible to bring about changes that can reduce crime. This is evident in policing today in terms of directed patrols to crime "hotspots" within cities. Using the crime data and other public data recorded by the city of Chicago, this team will search for trends in crime that perhaps could be used for social action.

University of Colorado at Boulder,

DOI: 10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

The existing content analyzed in this project often categorizes type of crimes into two groups: index and non-index crimes. Index crimes include homicides, criminal sexual assaults, aggravated assaults/batteries, burglaries, thefts, robberies, arson, and motor vehicle thefts. Non-index crimes include all other crimes, such as vandalism, weapons violations, public peace violations, and many others. Sources, such as the report by Andrew Papachristos describe the spatial trends of crime in Chicago as well as the changes in terms of frequency over time of index crimes and the overall crime rate. According to Papachristos' report, Chicago's crime rates are similar to those of other U.S. cities. Crime rose in the 1960s, reached a high point in the 1990s, and since then it has steadily declined (Papachristos 4). Additionally, the report mentions that "especially socially and economically disadvantaged communities continue to have stubbornly high levels of crime" (Papachristos 6). Given the data used in this project, the team should be to see this second trend in the rates of crime by community area in Chicago. Other sources found online present visualizations detailing the prevalence of types of crimes, such as theft versus battery. They also present time series plots of crime over the last 17 years. An additional point of study was the trends of a particular crime, such as sexual assault over a single year (Laughlin; Mangipudi).

3 DATA SET

The main data set comes from Chicago Data Portal, an online repository of public data managed by the city of Chicago. The data set comprises individual crime records from 2001 to February 2018. All together there is roughly 6.54 million crime records in this data set. Each record is comprised of 22 attributes:

- A unique identifier of the crime. This attribute was not used in the data mining process.
- A case number used by the Chicago Police Department. This attribute was also not used.
- The date the crime occurred. This attribute was used to group crimes for longitudinal study.
- The city block where the crime occurred. This attribute was used for grouping crimes within community areas.
- The IUCR crime code, which categorizes the type of crime. Using another data set available on the Chicago Data Portal,

the team was able to differentiate between index and non-index crimes in the dataset.
- The primary type of the crime, which is linked to the IUCR type, but it generalizes the type of crime into a categorical string. This attribute was used frequently along with the date field to find out how much crime occurred on a given day or within a date range.
- The description of the crime. This attribute gives more specific information about the crime than just the type of crime.
- A location description, which categorizes the type of location where the crime occurred (e.g. porch, sidewalk, apartment).
- A boolean field indicating whether an arrest was made for the crime.
- Another boolean field indicating whether the crime was domestic as defined by the Illinois Domestic Violence Act.
- The community area where the crime was committed. Chicago is comprised of 77 community areas. This attribute was similarly to the block attribute to group crimes by geographical areas.
- There are other interesting attributes in the dataset, but they were not used in this project.

This data set was the main one used in this project, but several others were used. Another data set used in this project was information on Chicago's affordable housing, which lists Chicago's affordable housing units by community area. A third data set provides information on access to prenatal care by trimester from 1999 to 2009 (also grouped by community area). A fourth dataset details the rates of elevated levels of lead in blood by community area from 1999 to 2013. The last data set used in this project was a census data set produced by the city of Chicago with various socioeconomic indicators for each of Chicago's 77 community areas, such as: overcrowded housing, unemployment, per capita income, and other figures.

4 TOOLS:

4.1 Python:

Ehe majority of the code written for this project was in Python. Many data discovery and analysis packages for Python that work well with large data sets like the one used in this project. In addition, the language is easy to prototype for quick data exploration.

4.2 Jupyter NoteBooks:

Jupyter Notebooks provide an easy and intuitive way to create interactive documents that work well for data presentation. These notebooks allow the user to both write Python code and Markdown. This allowed us to perform data analysis and keep notes of our findings along the way.

4.3 Pandas:

Pandas was essential to this project. It allowed the team to slice and manipulate the data set. In addition, it provides tools for data cleaning processing. One of the tasks we used Pandas frequently for was aggregating and grouping. For example, it allowed us to group crime records by day or month and aggregate the total amount of crime that was committed in that time period.

4.4 Numpy:

Numpy is a Python library for scientific computation with an emphasis on vector and matrix operations. The team used numpy to produce a polynomial regression model on the rates of crime over time.

4.5 Folium:

Folium is a great tool for geographically mapping data. We used it to heat-map the crime locations in Chicago. Although it is a great tool for seeing how the crime-rate moved to different areas of Chicago, because of the way the latitude and longitude are defined by the city block in the dataset, we weren't able to zoom in and see any significant change in any individual neighborhood.

4.6 Matplotlib:

Matplotlib provides an easy to use interface for constructing graphs and plots in Python. Some of the plots we created using Matplotlib include: histograms, scatter plots, and line graphs.

4.7 Scipy:

Scipy provides similar tools to Numpy in terms of scientific computation. It provides packages for statistical modeling as well as probability distributions. Specifically, we used Scipy for linear regression on crime rates over time.

4.8 RapidMiner:

RapidMiner was used for K-Means clustering. It allowed us to cluster the communities by their total number of crimes and plot all the data. It was fairly easy to create a flowchart (Figure 1) to manipulate the data use all of the tools.

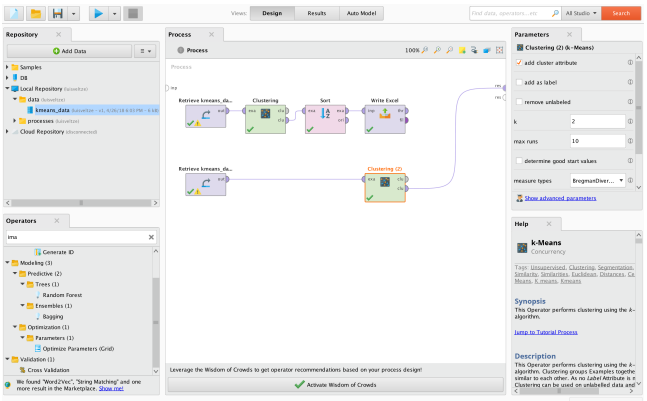

Figure 1: *RapidMiner Desgin*

4.9 Seaborn:

Seaborn is a great tool for creating nice looking data visualizations. We used it to create some of the more complicated plots that we used in our presentations.

5 MAIN TECHNIQUES APPLIED

The first step in this project was data cleaning and preprocessing. Surprisingly, not much data cleaning was necessary on this data set. The categorical attributes, such as the the crime type, location type, and others were consistent values. The team also did not have to deal with missing values in the data set. Most of the data set storage and manipulation was managed using Python's Pandas library. Given that the data set was already formatted correctly, much of the processing was just telling Pandas what type of data type a column should be in a dataframe.

One of the tasks of preprocessing was to reduce the data set to a manageable size. On disk, the data set is about 1.5 Gigabytes which makes computationally expensive operations very slow. For some data mining tasks it was necessary to use all six and a half million rows, but for other operations it was helpful to use a data set that represented an aggregation of the larger one and could be loaded into memory on a laptop. In order to do this, Pandas was used to break up the main data set into 18 new data sets grouped by year. The records in these datasets describe how many total crimes, index crimes, and crimes by type that were committed each data per community area in Chicago. Aggregating the data by day and community area reduced the entire dataset to about 415,000 records, which was much easier to manage.

Once this was done, these yearly data sets were used to plot the amount of crime committed each day of the year (Figure 2). Additionally, the crime by type, such as battery or theft, were plotted for each year (Figure 3). Examining these plots, it seems that the city of Chicago did not record all crime from the beginning of 2001 until around March of 2002. This is apparent in the drastic change in the plot of 2002 crime as it jumped from about 150 crimes in one day before March to around 1,300 per day in April. Another feature than can be observed from these graphs is that the rate in total crime and index crime follows a fairly consistent pattern. The shape of the scatter plot is slightly bell-shaped where the number of crime is highest roughly May through July. Interestingly, the total crimes spikes in early January especially around New Year's Day. This might be due to data collection issues on the part of Chicago's Police Department. Perhaps, if they were not sure when a crime occured in a year they rounded up or rounded down to the first of January.

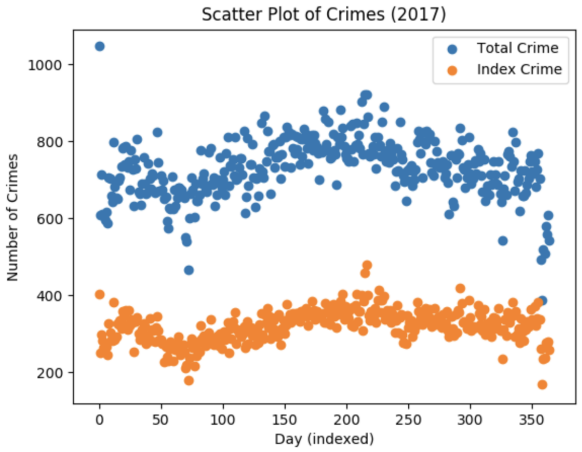

Figure 2: *Scatter Plot of Crimes*

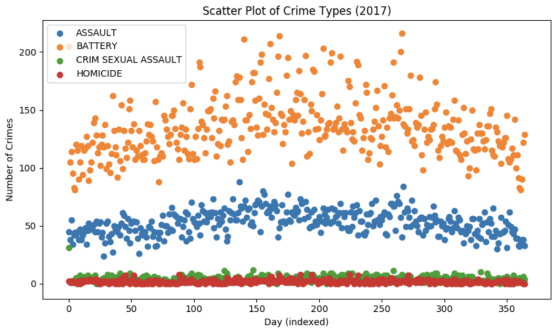

Figure 3: *Scatter Plot of Crime Types*

The yearly data sets were combined into a single dataframe for crime across all 18 years studied. From this graph (Figure 4) it is easy to see that overall the total number of crimes committed in Chicago steadily declined from 2003 to present-day 2018. The years 2001 and 2002 were excluded from this plot because of the underreporting issues for this time range as mentioned above. In addition, a simple linear regression was was run on the dataframe using Python Scipy and it too revealed a declining rate (a negative slope). Numpy's polynomial regression was also run on the data. It performed better than

the linear methods because it better fit the month to month fluctuations. This method also showed a decreasing amount of total crime.

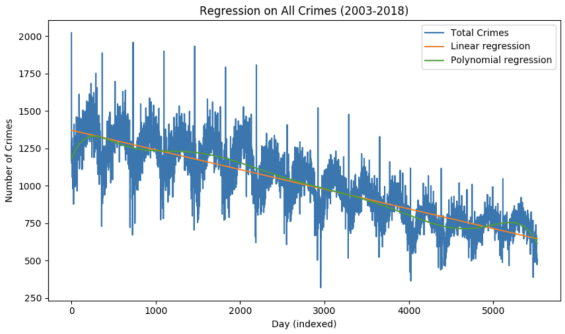

Figure 4: *Regression on all Crimes*

A third task was finding the outliers in the crime data. This, like the plots described above, was performed year to year and then over the entire time range of the data set. A subset of crime types was examined including: assault, battery, criminal sexual assault, homicide, stalking, and prostitution. These crimes were chosen out of pure curiosity, but further analysis should investigate all crime types. For each year, the program found the mean and standard deviation in the number of crimes committed of that type. Outliers were identified by looking for days in which the number of crimes was above or below a certain number of deviations. In this case, the team used three standard deviations and only looked at the days where the number of crimes exceeded this amount. The program found days in the year that matched this criteria and logged it to a file. In addition, the number of total crimes of that type was summed up and plotted across the year. Several interesting trends were seen from this data. Firstly, many of the outliers are around the middle of the year which reflects the overall yearly trend described above. Secondly, crimes of theft are particularly high on the first of the month. Secondly, homicides are higher during the early June to late September period. Thirdly, cases of battery and criminal sexual assault are particularly high around New Year's. In addition, the program determines which community areas are outliers in terms of index crime each month. Several community areas stood out as outliers: 8 and 25, which correspond to Near North Side and Austin respectively. Further work on outlier detection needs to take into account the population density of each community area over time.

One of the goals described in the project proposal was to incorporate multiple datasets in this project. These datasets would describe additional information about Chicago including census data, social data, and public health. Four datasets from the city of Chicago's online data catalog were added to yearly dataframe. The first dataset listed the number of affordable housing units per community area in Chicago. The second dataset provided information on elevated blood level levels in children ages 0-6 between the years of 1999-2013. The third dataset gave data on rates of prenatal care by community region in Chicago from 1999 to 2009. The fourth dataset listed static socioeconomic figures for each community region, including information about overcrowded housing, households below the poverty line, the percent unemployed or without a high school diploma, and per capita income. The crime dataframe was then grouped by community area in order to get the total number of crimes committed in a community per year. Using these yearly crime totals and social metrics from the other datasets, a correlation matrix was created using Pandas. From this matrix it appears that there is a slight positive correlation in the amount of crime and the rate where no prenatal care was given. Similarly, there were slight positive correlations with features of elevated blood lead levels, percent below poverty line, and percent unemployed. However, this matrix needs to be analyzed warily for a couple reasons. Firstly, the social data used was either static or did not span the same years of the crime dataset (2001-2018). In order to handle this, the missing values were filled with the average value in that column. This of course skews the data because it does not take into account some of the community changes not described in the time range. For example, the per capita income of each of Chicago's communities probably did not remain static from 2001 to 2018. Secondly, the correlation needs to take into more granular data about each community's crime statistics instead of just overall crime rates. Despite this, the matrix displays other interesting data such as the correlation between prenatal care rates and poverty or unemployment rates and blood lead levels.

In addition to the work described above other tasks were worked on. One of the tasks attempted was to find a correlation between the day of the year, the community region, and the number of a type of crime committed. However, this correlation matrix showed only weak support the way it was constructed. Alternative implementations should be explored to discover the correlations between community, time of the year, and crimes rates. Another task attempted was to predict the community region given crime record that lists the type of crime, the location type, whether a an arrest was performed, whether it was a domestic crime, and the year it was committed. ScikitLearn was used to train a multiclass random forest classifier to identify the community region. This process was run on all 6.5 million rows of the dataset and after 11 hours of running it showed no signs of stopping. Future investigation should try random sampling to get a smaller subset of the entire dataset for training. A third task, which is currently in development, is to predict the number of crimes of one type given the number of crimes of another type. For example, given the number of crimes of narcotics, robbery, theft, burglary, stalking, and assault, can you predict the number of sexual assault crimes that will occur on that day and in a specific community?

In an attempt to better understand the data set we created some visualizations in order to gain some additional insight. Some of the most useful visualizations that we created were heat-maps generated using the Python library Folium. Fortunately, Folium works well with Pandas which allowed us to easily create heat-maps of different types of crimes (Figure 5). One particularly interesting trend that we noticed when we were comparing different heat-maps was the drastic change in the number of prostitution arrests between the years 2001-2017. Another trend that we spotted while doing visualizations was the drastic change

in crime rate in the 100xx block of W O’Hare Street. Which, in the early 2000’s had one of the worst crime rates in Chicago. But, after a few years, the crime rate in that area decreased drastically. We were wondering if there was a simple reason for this sudden decline. We did some research and discovered that the City of Chicago had invested a significant amount of money in efforts to revitalize the area. Who would have thought that there was such a simple way to decrease the crime rate in an area, invest some time and money in making it nicer.

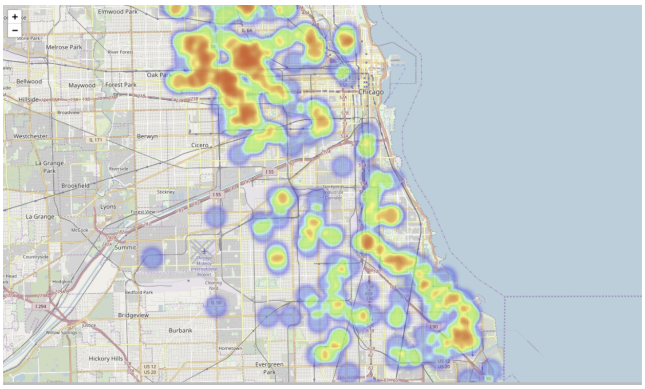

Figure 5: ***Heat-map of Crimes in Chicago***

One particularly interesting trend that we noticed when we were comparing different heat-maps was the drastic change in the number of prostitution arrests between the years 2001-2017(Figure 6 - Figure 7).

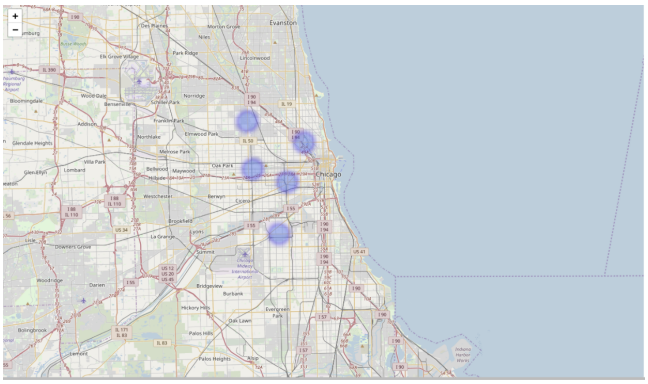

Figure 6: ***Heat-map of Crimes in Chicago, 2001***

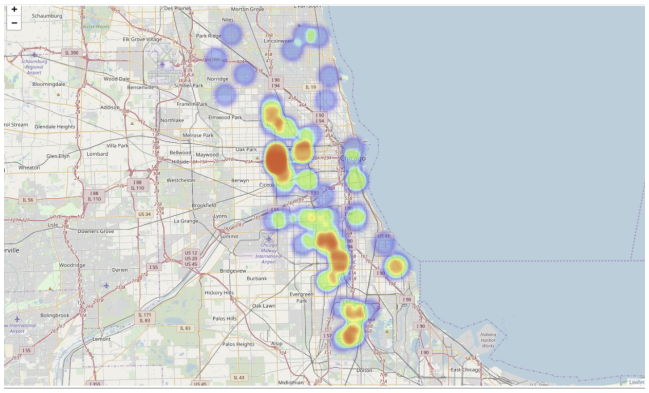

Figure 7: ***Heat-map of Crimes in Chicago, 2017***

The final data mining technique that the team used in this project is K-means clustering. In order to apply this technique, we needed to create a new data set with only the community areas, total crimes per year, and sum of all crimes across the 17 years. This new dataset set was created in Jupyter Lab using the Python libraries Pandas, Matplot, and Numpy. We then tried to do the clustering directly on Python using Sklearn library but we were unsuccessful at doing so. We were unsuccessful because even though we were able to generate three centroids (Figure 8), the data points were not clustering to a centroid. After many failed attempts at coding k-means clustering in python, we ended up using a data mining software called RapidMiner studio.

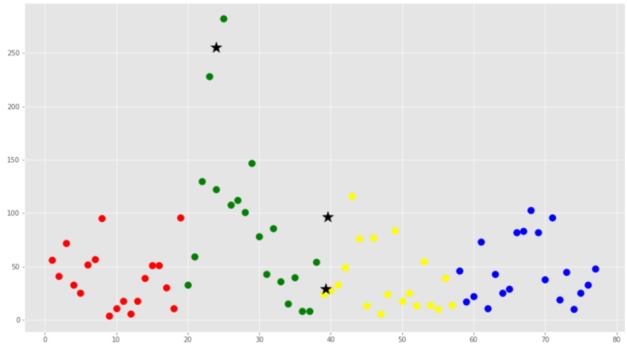

Figure 8: ***K-Means clustering in Python***

We first had to import the k-means data file to RapidMiner. Then we used some of RapidMiner’s tools to cluster the data and export the clustered data set as well as to plot the clusters. A pie chart (Figure 9) and a scatter plot (Figure 10) were generated to represent the clustered dataset. The pie chart consisted of three pies, red, blue, and green, in decreasing number of crimes respectively. Finally, we colored a Chicago map (Figure 11), that was divided by communities, to show which communities belonged to which cluster.

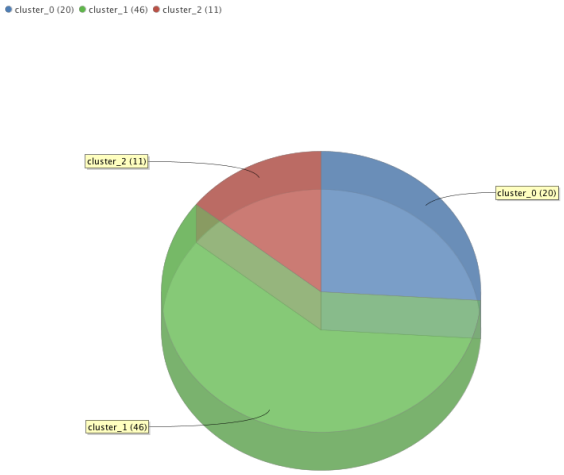

Figure 9: ***K-Means Clustered Pie Chart***

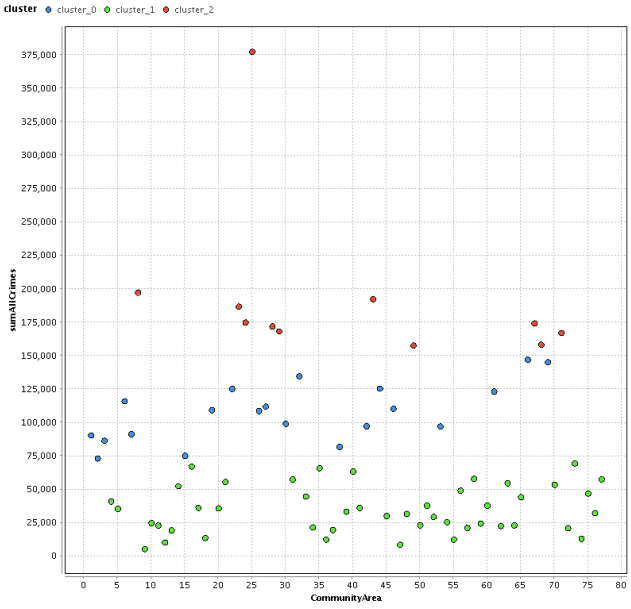

Figure 10: ***K-Means Scatter Plot***

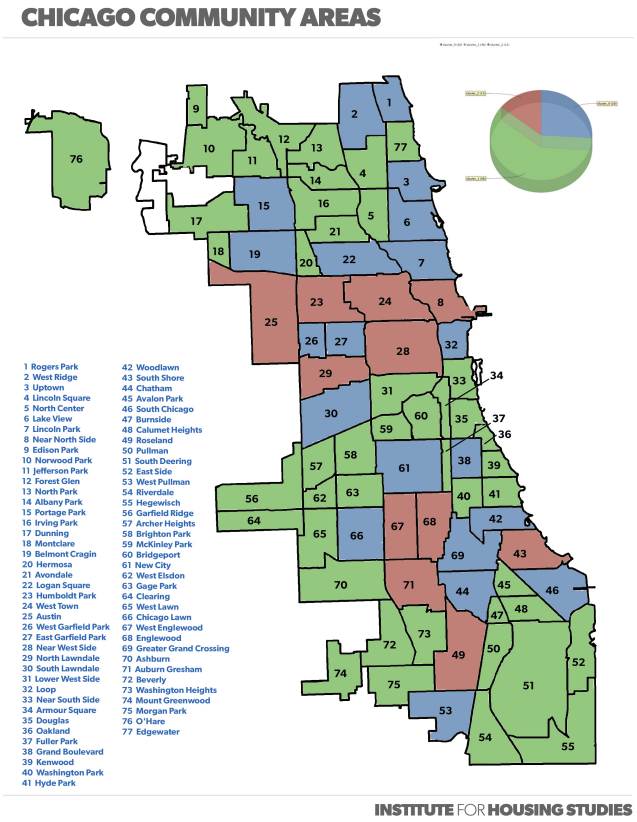

Figure 11: ***Chicago Community Clusters for Number of Crimes***

6 KEY RESULTS

The team learned many interesting things in the process of this project. One of the things that should be noted is that given the nature of the topic the team chose, crime data, it is difficult to draw concrete conclusions because there are so many variables involved. In spite of this, there are several notable features we found in the data.

Firstly, the overall trend of crime in Chicago is downwards. Over the roughly 17 years studied in the project the total crime has decreased since 2001. It is a different story for different types of crimes, where we see more variation. For example, homicides rose during this time range, while criminal sexual assault declined. Other crimes, such as theft and battery, decreased but at a much more smaller rate. This information was discovered by running a linear regression on a dataset of the number of crimes of that type per day for the the 17 plus years. It is also notable that the amount of that type of crime varies by the time of year. For example, crimes such as theft, assault, and battery are generally higher during the summer months.

Through outlier detection, several interesting trends where found. Firstly, crimes of battery are particularly high in June. Many homicides occur in the months of July and August. Theft and sexual assault crimes show an interesting pattern too: the outliers often occur on the first of the month. According to Chicago's data catalog, the date field is the day when the crime occurred or "a best estimate." It is possible that for these types of crimes there is more uncertainty about the day that it was committed. Perhaps, the victim only realizes days later something was taken from them, but they do not know the specific day it happened so the police round to the first of the month. There are, of course, other possible reasons that the specific date was not reported correctly. However, if we can trust these dates or at least trust that they are good estimates of the true value then these trends are an important piece of knowledge. According to the program, sexual assault spikes on the first of January. For example, in 2015 the average number of sexual assaults per day is 3.75, but on January first there are 52! The second part of outlier detection was finding community areas that stood out with particularly high rates of Index crime. This was done on a month to month basis. Community areas 8 and 25, which correspond to Near North Side and Austin, stood out frequently as regions with high rates of Index crime. However, other communities such appeared as outliers but only during certain months in the year. For instance, in 2016 the community region of Loop was an outlier at the beginning and end of the year. In 2017, Loop was an outlier for the second of the year. This might mean that index crime is becoming more common in Loop.

The work done to connect crime data to the socioeconomic and health data revealed some interesting numbers, but unfortunately they are more difficult to interpret. As described above, the crime rates were taken on a yearly basis and then correlated with data on various social topics, such as access to prenatal care and affordable housing as well as metrics on poverty line and unemployment. One of the findings revealed a slight positive correlation between the lack of prenatal care (in any trimester) and the crime rate. However, when there was prenatal care in the first trimester there was a slight negative correlation with the crime rate. Similarly, there is a slight positive correlation between crime rates and the rates of elevated blood lead levels. The correlation matrix also shows positive correlations between crime rates and high rates of overcrowding, percentage below the poverty line, and unemployment. Although not technically relevant to our project, there other interesting correlations found in this process. For instance, there is a correlation between lack of prenatal care and elevated blood lead levels and rates of poverty. Overall, these findings are similar to those discussed in the Related Work section: socially disadvantaged communities experience higher rates of crime. There is a caveat with this correlation data because the data used in the correlations in some cases needed to be extrapolated to match the years of study. In addition, it is important to remember these variables do not represent causality for rates of crime. Instead, this represents simple correlations between rates of crime and a given variable. There are many social, economic, and health factors that go into crime.

Our most successful AI algorithm was K-means, it gave us three definitive clusters. Cluster 1 was comprised of 11 community groups, this cluster is the one associated with high crime rates, it can be seen as red in figure 11. The second cluster that was identified was comprised as moderate crime rate and had 20 members, it can be seen as the blue shaded regions in figure 11. Finally, cluster 3 which had 46 members, was identified as being a low crime area, it can be seen as green in figure 11. One of the main trends we noticed is that over time crime has been slowly concentrating to specific areas. Our clustering algorithm re-enforced our initial hypothesis that crime was centralized to different pickets. Overall, our main take away from this clustering algorithm is that crime has become condensed to two pockets of the city. The highest crime communities are bordered by moderate crime communities and whose pockets of high and moderate crime are completely surrounded by low crime areas.

7 APPLICATIONS

The knowledge gain in this project has important social value. From what we have found, there are four major factors to rates of crime in Chicago: geography, time, public health, and economy. This information is valuable to law enforcement and lawmakers that can respond to these factors. For example, we know that certain communities, even neighborhoods, with higher rates of poverty or poor public health will experience higher rates of crime especially during the summer months. The Chicago Police Department could use the information found by this project to target sexual assault crimes which were found to happen frequently on the first of the year. Or they could increase directed patrols on the first of the month because it was found that thefts are more frequent on the first of the month. Similarly, it might be useful for the Chicago Police Department to consider shifting resources from communities where crime is less serious to areas where crime rates are higher. For lawmakers, the knowledge gained in this project could help guide funding to areas with poor public health or economic situations.

By obtaining a better understanding of which factors contribute toward the crime rate of an area, we can better understand what needs to be done to prevent future crimes. Although it may seem obvious to some that people in poverty-stricken neighborhoods are more likely to be exposed to criminal elements. It may not be obvious what steps can be taken to limit this. Increasing funding toward community improvements in less-fortunate areas of Chicago can have a long lasting effect on the overall crime-rate of the area, as it did on W O'Hare Street. By increasing available funding for pregnant mothers to receive adequate pre-natal care in affordable housing neighborhoods would likely have a positive impact on the crime in the area. Knowing that, as time progresses, the crime-ridden areas of Chicago are becoming more condensed and isolated will help to better address the situation in the future.

Unfortunately, none of these applications are necessarily easy to implement. It takes time to acquire the funding needed to make a change. A lot of people have to contribute if we want to see improvement in the near future. Convincing lawmakers to change their approach to handling crime-ridden areas could prove

to be near impossible. But by analyzing the data and attempting to understand what changes might actually help with the crime in Chicago is an important first step.

8 WORKS CITED:

Mangipudi, Vivek. ANALYSIS OF CRIMES IN CHICAGO 2001 - 2017. 28 July 2017, rstudio-pubs-static.s3.amazonaws.com/294927_b60231 8d06b74e4cb2e6be336522e94e.html

Papachristos, Andrew V. "48 YEARS OF CRIME IN CHICAGO: A Descriptive Analysis of Serious Crime Trends from 1965 to 2013." Yale ISPS, vol. 13, no. 023, 9 Dec. 2013, pp. 1fi?!20., isps.yale.edu/sites/default/ files/publication/2013/12/48yearsofcrime_final_ispsworkingpaper023. pdf

Laughlin, Greg. "Crime Over Time: Visualizing Crime Data in Chicago." Socrata, 3 June 2014, socrata.com/blog/crime-time-visualizing-crime-data-chicago/.